Cognition 120 (2011) 380–390

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis

Anne S. Hsu [a,*], Nick Chater [b], Paul M.B. Vitányi [c]

[a] Department of Cognitive, Perceptual and Brain Sciences, University College London, 26 Bedford Way, London, WC1H 0AP, UK
[b] Behavioural Science Group, Warwick Business School, University of Warwick, Coventry, CV4 7AL, UK
[c] National Research Center for Mathematics and Computer Science in the Netherlands (CWI), Science Park 123, 1098 XG Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Available online 26 March 2011

Keywords:
Child language acquisition
Poverty of the stimulus
No negative evidence
Bayesian models
Minimum description length
Simplicity principle
Natural language
Probabilistic models
Identification in the limit

ABSTRACT

There is much debate over the degree to which language learning is governed by innate language-specific biases, or acquired through cognition-general principles. Here we examine the probabilistic language acquisition hypothesis on three levels: We outline a novel theoretical result showing that it is possible to learn the exact *generative model* underlying a wide class of languages, purely from observing samples of the language. We then describe a recently proposed practical framework, which quantifies natural language learnability, allowing specific learnability predictions to be made for the first time. In previous work, this framework was used to make learnability predictions for a wide variety of linguistic constructions, for which learnability has been much debated. Here, we present a new experiment which tests these learnability predictions. We find that our experimental results support the possibility that these linguistic constructions are acquired probabilistically from cognition-general principles.

1. Introduction

Much research suggests children can learn language from mere exposure, without relying on other's feedback about their own utterances, i.e., from positive evidence alone. How children thus learn language has been a heavily researched topic. Two main perspectives on language acquisition can be understood through the distinction between *discriminative* and *generative* learning models (Hsu & Griffiths, 2009). A discriminative model learns by establishing a boundary between categories by mapping inputs to categories from a set of input-category pairs. For language, these are categories of grammatical and ungrammatical sentences. From the discriminative perspective, the ability to learn from only positive examples seems puzzling: with only positive examples, i.e., grammatical sentences, a discriminative learner has no basis on which to determine the boundary between grammatical and ungrammatical sentences. Under discriminative based perspectives, theoretical analyses of learnability from only positive examples tend to be discouraging (Gold, 1967; Nowak, Komarova, & Niyogi, 2002). Indeed, thinking about language learning as a classification problem has led many theorists to conclude that language acquisition faces fundamental "logical" problems (Baker & McCarthy, 1981; Horning, 1969).

The Bayesian approach to cognitive development, explored in this special issue, and the cognitive sciences more generally, e.g., Griffiths, Chater, Kemp, Perfors, and Tenenbaum (2010), suggest a different perspective on learning: generative models learn by making inferences about the probability distribution that produces the language input. Thus, from a generative perspective, language acquisition is not a matter of discriminating "good" from "bad" linguistic forms; instead the aim is to model the underlying regularities that give rise to the language. The key assumption

* Corresponding author.
E-mail addresses: ahsu@gatsby.ucl.ac.uk (A.S. Hsu), n.chater@ucl.ac.uk (N. Chater), paul.vitanyi@cwi.nl (P.M.B. Vitányi).

0010-0277/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.02.013

Table 1
Sentences used in experiment: quadruplets illustrating restriction rule.

	Set 1	Set 2
	(a) restricted form of construction with restriction*	(a) restricted form of construction with restriction*
	(b) un-restricted form of construction with restriction	(b) un-restricted form of construction with restriction
	(c) restricted form of related construction	(c) restricted form of related construction
	(d) un-restricted form of related construction	(d) un-restricted form of related construction
1	(a) I think it is fine where it's.*	(a) Do you know who she's?*
	(b) I think it is fine where it is.	(b) Do you know who she is?
	(c) I think it's fine for now.	(c) Do you know who she's waiting for?
	(d) I think it is fine for now.	(d) Do you know who she is waiting for?
2	(a) He arrived the plane safely.*	(a) The captain arrived the ship on a small island.*
	(b) The plane arrived safely.	(b) The ship arrived on a small island.
	(c) He landed the plane safely.	(c) The captain landed the ship on a small island.
	(d) The plane landed safely.	(d) The ship landed on a small island.
3	(a) I am glad he came the helicopter in time.*	(a) The pilot came the plane ten minutes ahead of schedule.*
	(b) I am glad the helicopter came in time.	(b) The plane came ten minutes ahead of schedule.
	(c) I am glad he landed the helicopter in time.	(c) The pilot landed the plane ten minutes ahead of schedule.
	(d) I am glad the helicopter landed in time.	(d) The plane landed ten minutes ahead of schedule.
4	(a) Tom donated the hospital a large amount of money.*	(a) Charles donated the library some very valuable books.*
	(b) Tom donated a large amount of money to the hospital.	(b) Charles donated some very valuable books to the library.
	(c) Tom gave the hospital a large amount of money.	(c) Charles gave the library some very valuable books.
	(d) Tom gave a large amount of money to the hospital.	(d) Charles gave some very valuable books to the library.
5	(a) He fell the crumbs on the ground.*	(a) During the earthquake, he fell the wine glass.*
	(b) The crumbs fell on the ground.	(b) During the earthquake, the wine glass fell from his hands.
	(c) He dropped the crumbs on the ground.	(c) During the earthquake, he dropped the wine glass.
	(d) The crumbs dropped on the ground.	(d) During the earthquake, the wine glass dropped from his hands.
6	(a) She disappeared her money behind the curtains.*	(a) Cathy was able to disappear her anger.*
	(b) She disappeared behind the curtains.	(b) Cathy was able to disappear.
	(c) She hid behind the curtains.	(c) Cathy was able to hide her anger.
	(d) She hid her money behind the curtains.	(d) Cathy was able to hide.
7	(a) That is an unusual object What's it?*	(a) I know something is bothering you. What's it?*
	(b) That is an unusual object. What is it?	(b) I know something is bothering you. What is it?
	(c) That is an unusual object. What's it used for?	(c) I know something is bothering you. What's wrong?
	(d) That is an unusual object. What is it used for?	(d) I know something is bothering you. What is wrong?
8	(a) I poured the truck with gravel.*	(a) Kate poured the fish tank with pebbles.*
	(b) I poured the gravel into the truck.	(b) Kate poured the pebbles into the fish tank.
	(c) I loaded the truck with gravel.	(c) Kate loaded the fish tank with pebbles.
	(d) I loaded the gravel into the truck.	(d) Kate loaded the pebbles into the fish tank.
9	(a) He vanished the treasure inside a cave.*	(a) Ben vanished his toys behind the door.*
	(b) He vanished inside a cave.	(b) Ben vanished behind the door.
	(c) He hid the treasure inside a cave.	(c) Ben hid his toys behind the door.
	(d) He hid inside a cave.	(d) Ben hid behind the door.
10	(a) Susan created her daughter a new dress.*	(a) The chef will create you a special dish.*
	(b) Susan created a new dress for her daughter.	(b) The chef will create a special dish for you.
	(c) Susan made her daughter a new dress.	(c) The chef will make you a special dish.
	(d) Susan made a new dress for her daughter.	(d) The chef will make a special dish for you.
11	(a) Someone is at the door. Who's it?*	(a) I heard you found a date for the party. Who's it?*
	(b) Someone is at the door. Who's there?	(b) I heard you found a date for the party. Who's the lucky girl?
	(c) Someone is at the door. Who is it?	(c) I heard you found a date for the party. Who is it?
	(d) Someone is at the door. Who is there?	(d) I heard you found a date for the party. Who is the lucky girl?
12	(a) Dan is gonna a picnic in the park.*	(a) My sister is gonna Disneyland for her school trip.*
	(b) Dan is going to a picnic in the park.	(b) My sister is going to Disneyland for her school trip.
	(c) Dan is gonna attend a picnic in the park.	(c) My sister is gonna be at Disneyland for her school trip.
	(d) Dan is going to attend a picnic in the park.	d) My sister is going to be at Disneyland for her school trip.
13	a) James suggested his manager the plan.*	a) Rachel suggested the client her new idea.*
	b) James suggested the plan to his manager.	b) Rachel suggested her new idea to the client.
	c) James told his manager the plan.	c) Rachel told the client her new idea.
	d) James told the plan to his manager.	(d) Rachel told her new idea to the client.
14	(a) That is an unusual object. What's it?*	(a) Who do you think that is meeting with the boss?*
	(b) That is an unusual object. What is it?	(b) Who do you think is meeting with the boss?
	(c) That is an unusual object. What's it used for?	(c) Who do you think that the boss is meeting with?
	(d) That is an unusual object. What is it used for?	(d) Who do you think the boss is meeting with?
15	(a) Which player does does Richard wanna win?*	(a) Who do you wanna win the championships?*

(*continued on next page*)

Table 1 (*continued*)

	Set 1	Set 2
	(b) Which player does Richard want to win?	(b) Who do you want to win the championships?
	(c) Which player does Richard wanna beat?	(c) Who do you wanna beat in the championships?
	(d) Which player does Richard want to beat?	(d) Who do you want to beat in the championships?
16	(a) I'd rather the spaghetti.*	(a) I would rather to see the later show.*
	(b) I'd prefer the spaghetti.	(b) I would prefer to see the later show.
	(c) I would rather have the spaghetti.	(c) I would rather the later show.
	(d) I would prefer the spaghetti.	(d) I would prefer the later show.

for generative models is that the input is sampled from the natural language distribution; discriminative models do not require this assumption.

In this paper, we present a generative Bayesian perspective on the problem of language acquisition spanning three levels of analyses, theoretical, computational and experimental. The theoretical and experimental results are novel contributions of this paper and the computational results are summarized from recent related work. First, we present new theoretical learnability results, indicating that under fairly general conditions, it is possible to precisely identify the generative language model. Combined with prior work, these results suggest that the "logical" problem of language acquisition may be reduced by adopting a probabilistic perspective. Second, we review a recently proposed, general framework which quantifies Bayesian learnability of constructions in natural language under a simplicity principle. This framework has been used to predict natural language learnability for a wide variety of linguistic rules, using corpus data. Third, we present a new experiment which tests these learnability predictions by comparing them with adult grammaticality judgments.

2. Gold revisited: generative model identification in the limit

A central theoretical question is: given sufficient exposure to the language, can the learner recover a perfectly accurate description of that language? Gold (1967) showed under certain assumptions this is not possible. However, a range of more positive results have since been derived, e.g., (Angluin, 1988; Chater & Vitányi, 2007; Feldman, Gips, Horning, & Reder, 1969; Horning, 1969). These results apply across linguistic levels: including the acquisition of phonology, morphology, syntax, or the mapping between syntax and logical form. Crucially, our results rely on the assumption of the *computability* of the probability distribution from which language is sampled. This assumption follows naturally from a computational perspective on the mind, and hence language production: but turns out to radically simplify the problem of language acquisition.

Specifically, we outline a new and strong positive learnability result. Our most basic result is the *Computable Probability Identification Theorem*. Although the result applies more generally, we will frame the discussion in terms of language. Informally, the theorem can be stated as follows: suppose an indefinitely long language corpus is generated by identical independently distributed (i.i.d.) samples from some probability distribution, p, over a countable set (e.g., the set of sentences in a language). We require only that p is computable 1[1]: there exists some computational device (e.g., a Turing machine), that, for each x, can compute the probability $p(x)$.

Then there exists a computable algorithm that receives the sentences, in sequence, and generates a sequence of guesses concerning the generative probabilistic model of the language (see Appendix). We will call these guesses $q_1, q_2, \ldots$. After sufficient linguistic data, the algorithm will almost surely alight on a guess which subsequently never changes: that is the sequence $q_1, q_2, \ldots$ "converges" almost surely to q. Moreover $q = p$, the probability distribution generating the language. This implies that the learner can then itself *generate* language using the correct probability distribution.

This result indicates not only that there need not be a "logical" problem of language acquisition (Baker & McCarthy, 1981; Hornstein & Lightfoot, 1981); but provides an algorithm which defines a computable process that will almost surely define (precisely if the sequence of samples is typical) not only the language, but the precise generative probabilistic model from which the language.

This result is stronger than many previous results, in a number of ways. (1) The language can be learned from the entire class of computable generative probability distributions for language. Thus, the method is not restricted to particular classes of language structure, such as finite state or probabilistic context-free grammars, in contrast, to many important theoretical results (e.g., Angluin, 1988; Clark & Eyraud, 2007; Feldman et al., 1969). (2) The learner does not merely approximate the language, as in most positive probabilistic results, but identifies the generating distribution precisely. (3) The learning method is computable (in contrast, for example, to Chater & Vitányi, 2007).

A number of open questions remain. The proof in the i.i.d. case depends on the strong law of large numbers. The question remains whether our results hold under weakened i.i.d. assumptions, e.g., sentences which have only "local" dependencies (relative to the total corpus the learner has been exposed to). In reality, there are complex interdependencies between sentences at many scales. These may arise from local discourse phenomena such as

[1] This is a mild restriction, which presumably holds for any cognitively plausible model of language production or mental representation of language; and clearly holds for standard probabilistic models of language, such as probabilistic context free grammar, n-gram models, or hidden Markov models (provided the parameters are computable).

anaphora, to much higher-level dependencies determined by the over-all coherence of a conversation or narrative. One possibility is that these dependencies will "wash out" over long time horizons. Additionally, for probabilistic processes which are stationary and ergodic, there are limit theorems analogous to the strong law of large numbers, raising the possibility that analogous results apply. Note that the present result is much stronger than traditional language identification in the limit (e.g., Osherson, Stob, & Weinstein, 1985): we show that the precise probability distribution generating language can be precisely identified (almost surely), not merely the set of sentences allowed in the language. A second open question concerns the impact of errorful linguistic input: can the learner infer the underlying "correct" structure, nonetheless? A third open question concerns the number of sentences typically required for identification to occur. We leave these, and other, open questions for a later technical paper (Vitányi & Chater, in preparation).

The present result might appear remarkably strong. After all, it is not generally possible to precisely identify the probability p with which a biased coin lands heads from a finite set of coin flips, however long.[2] The same applies to typical statistical language models, such as probabilistic context-free phrase structure grammars. These language models are typically viewed as having real valued parameters, which is a psychologically and computationally unrealistic idealisation that makes the problem of generative model identification unnecessarily difficult. In practice, any computational process (e.g., inside the head of the parent, to whom the child is listening) can only be determined by *computable* processes—and hence computable parameters, dramatically reducing the possible parameter values. We do not mean to suggest that the child can or does precisely reproduce the generative probabilistic model used by adult speakers. But if such identification is possible for any computable linguistic structure, the child presumably faces no insurmountable logical problems in acquiring a language from positive data alone.

3. A practical framework for quantifying learnability

We have presented above a new and strong learnability result. But much debate concerning language acquisition concerns more specific questions, such as how children learn restrictions to general rules from only positive examples. Restrictions on the contraction of '*going to*' provide an illustrative example: '*I'm gonna leave*' is grammatical, whereas '*I'm gonna the store*' is ungrammatical. Table 1 shows quadruplets (sentences *a*–*d*) relevant to restrictions on a variety of constructions: *a* is the ungrammatical, restricted form of the construction (e.g., contraction of *going to* where *to* begins a prepositional phrase). *b* is the grammatical un-restricted form (e.g., un-contracted form of *going to* is always allowed). *c* and *d* contain an analogous construction to that in *a* and *b*, but for which there is no restriction (e.g., contracted and un-contracted forms of *going to* where *to* is part of an infinitive verb). Sentences *c* and *d* are the basis for over-generalization of the restricted form in *a*.

Language acquisition requires the speaker to generalize from previously heard input. Research indicates that many (perhaps most) children are rarely corrected when they produce an over-general, ungrammatical sentence (Bowerman, 1988). These observations evoke the question: how do children learn which over-generalizations are ungrammatical without explicitly being told? Many language acquisition researchers have traditionally claimed that such learning is impossible without the aid of innate language-specific knowledge (Chomsky, 1975; Crain, 1991; Pinker, 1989).

More recently, researchers have shown that statistical models *are* capable of learning such rules from positive evidence only (Dowman, in preparation; Foraker, Regier, Khetarpal, Perfors, & Tenenbaum, 2009; Grünwald, 1994; Perfors, Regier, & Tenenbaum, 2006; Regier & Gahl, 2004). These statistical models are based on a particular instantiation of Bayesian modelling in which languages are chosen based on the principle of simplicity.

Recently, a *general quantitative framework* has been proposed which can be used to assess the statistical learnability of any given *specific linguistic restriction* in the context of real language, using only positive evidence (Hsu & Chater, 2010). This framework built upon previous simplicity-based modelling approaches (Dowman, in preparation; Foraker et al., 2009; Perfors et al., 2006) to develop a method that is generally applicable to constructions in natural language. When using this framework there are two main assumptions: (1) The description of the grammatical rule to be learned. (2) The corpus approximating the learner's input. Given these two assumptions, the framework provides a method for quantifying learnability from language statistics. The framework allows for comparison of results which arise from varying these two main assumptions, providing a common forum for quantifying and discussing language learnability. This framework assumes an *ideal statistical learner* and thus provides an upper bound on learnability based on language statistics. However, measures of learnabity should give an indication for how relatively statistically learnable constructions are in reality.

3.1. The minimum description length hypothesis

Because this framework is detailed elsewhere (Hsu & Chater, 2010), we only provide a brief overview here. Learnability evaluations under simplicity are instantiated through the principle of minimum description length (MDL). MDL is a computational tool that quantifies the information available to an idealized (cognition-general) statistical learner (Jacob Feldman, 2000). When MDL is applied to language, grammars are represented as a set of rules, such as that of a probabilistic context-free grammar (PCFG) (Grünwald, 1994). An information-theoretic cost is assigned to encoding grammar rules as well as to encoding the language under those rules. MDL has formal relations to Bayesian probabilistic analysis, although we do not

[2] To see this, note that the number of real values on the interval [0, 1] is uncountable, whereas the number of guesses associated with any infinite sequence of coin flips is countable. Therefore, the probability any of these guesses is correct has measure 0 in the standard uniform measure on the real interval [0, 1].

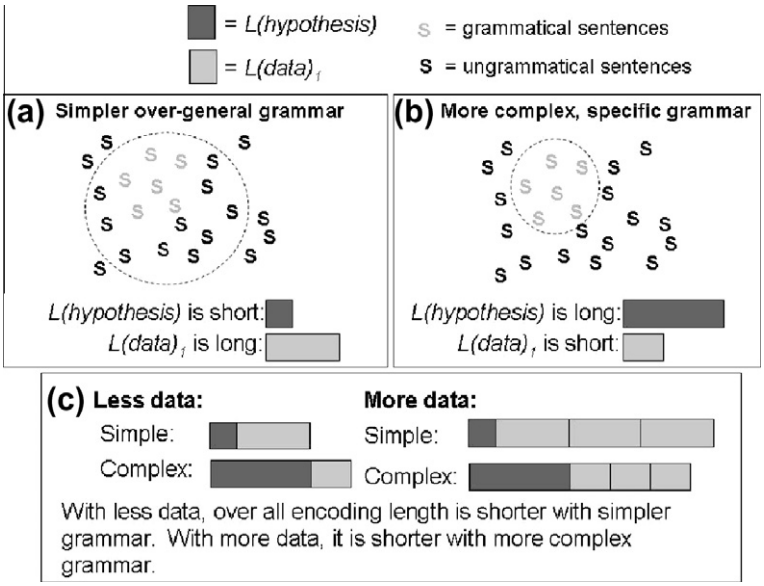

Fig. 1. *MDL simple grammar vs. efficient language encoding trade off.* (A) A simpler grammar is often over-general, i.e., allows for ungrammatical sentences as well as grammatical ones. Such an over-general grammar may be easy to describe (i.e., short grammar encoding length), but results in less efficient (longer) encoding of the language data. (B) A more complex grammar may capture the language more accurately, i.e., allows only for grammatical sentences and doesn't allow for ungrammatical sentences. This more complex grammar may be more difficult to describe (i.e., longer grammar encoding length), but will provide a shorter encoding of language data. (C) Initially, with limited language data, the shorter grammar yields a shorter coding length over-all, and is preferred under MDL. However, with more language input data, the savings accumulated from having a more efficient encoding of language data correctly favour the more complex grammar.

focus on this here (see Chater, 1996 for an informal description; Vitányi & Li, 2000 for a detailed analysis).

Hsu and Chater (2010) used two-part MDL. In the context of language acquisition, the first part of MDL specifies probabilistic grammatical rules to define a generative probability distribution over linguistic constructions, which combine to form sentences. Note that these are not necessarily the true probabilities in the language, but are the probabilities as specified under the current hypothesized grammar. The second part of MDL uses the probabilistic grammar to encode all the sentences heard so far. MDL selects the grammar that minimizes the *total* code length (measured in bits) of both the grammatical description and the encoded language length.[3]

According to information theory, the most efficient code occurs when each data element is assigned a code of length equal to the smallest integer greater than or equal to $-\log_2(p_n)$ bits, where p_n is the probability of the nth element in the data. For our purposes, these elements are different grammar rules. The probabilities of these grammar rules are defined by the grammatical description in the first part of MDL. Because efficient encoding results from knowing the correct probabilities of occurrence, the more accurately the probabilities defined in the grammar match the actual probabilities in language, the more briefly the grammar will encode the sentences in the language.

Under MDL, learners prefer grammatical descriptions that provide the shortest two-part code for the data received so far. Savings occur because certain grammatical descriptions result in a more efficient (shorter) encoding of the language data. If there is little language data (i.e., a person has little language exposure), encoding detailed specification of the language in the first part of the code will not yield large enough savings in the second part of the code to be chosen. Instead, a "cheaper", simpler, grammar will be preferred. When there is more language data, investment in a more costly, complicated grammar becomes worthwhile (see Fig. 1). This characteristic of MDL learning can explain the early over-generalizations followed by retreat to the correct grammar that has been observed in children's speech (Bowerman, 1988).

3.2. A practical example

We provide a brief example of how the framework is used to assess learnability of the restriction on contraction of *going to* (see Hsu and Chater (2010) for details). MDL part 1 assesses the difference between original and new grammar lengths (grammar cost). Here the old grammar allows *going to* to contract under all circumstances. The new grammar will enumerate situations where contraction of *going to* is not allowed. The difference between grammar encoding lengths came from defining the specific situations where *going to* can and cannot contract, i.e., [contractable going to] = [going to] [verb] and [not-contractable going to] [going to] [a place]. Concepts within brackets were represented as a single encoded symbol: [going to] represents use of the words *going to* in a sentence and [verb] represents any verb

[3] The MDL framework can also be expressed as a corresponding Bayesian model with a particular prior (Chater, 1996; MacKay, 2003; Vitányi & Li, 2000). Here, code length of the model (i.e., grammar) and code length of data under the model (i.e., the encoded language) in MDL correspond to prior probabilities and likelihood terms respectively in the Bayesian framework.

and [a place] represents any destination one may go to. These formally correspond to the use of *to* as part of an infinitive verb, e.g., *I am going to stay* and the use of *to* as a preposition meaning *towards*, e.g., *I am going to school*.[4] The additional encoding length cost from the additional definitions can then be quantified in bits through MDL, see Hsu and Chater (2010).

The second part of MDL requires the evaluation of savings under the new grammar. This requires specifying the occurrence probabilities (estimated from a chosen corpus) of each sentence under original vs. new grammars. Under the original grammar, contractions were always allowed, and finite code lengths were required to encode whether contraction occurs in all situations. Under the new grammar, *going to* contraction never occurs when *to* is a preposition and thus 0 bits are required to encode contraction. By comparing the difference in encoding costs under the original vs. new grammars, we can calculate the savings accrued per occurrence of *going to* contracted in the infinitive form (which is the only one where contraction is allowed). The number of occurrences needed to learn the construction is obtained by determining the amount of occurrences needed so that savings becomes greater than or equal to the cost in grammar length difference.

In summary, learnability is affected by (1) Complexity of the rule to be learned. More complexity increases grammar cost and decreases learnability. (2) Frequency of the restricted vs. un-restricted forms of a construction in other similar lexical items, e.g., frequency of sentence type c vs. d in Table 1. Greater frequency increases encoding savings and increases learnability. For example, for assessing the dative restriction on donate: if *give* appeared mostly in the direct dative (restricted form for *donate*), one would more quickly notice the suspicious absence of the direct dative form for the verb *donate*. (3) Occurrence frequency of the construction whose restriction is to be learned, e.g., type *b* sentences, such as use of *going to* where *to* introduces a prepositional phrase. (1) and (2) determine how many occurrences are needed for learning and (3) (estimated from corpora serving as input) then will determine how many years it will take to accrue the number of occurrences needed.

4. Testing learnability predictions

Hsu and Chater (2010) used the above framework to predict learnability for linguistic rules whose learnability have been commonly debated in the language acquisition field. These rules all involve restriction rules for the following 15 constructions[5]: contractions of *want to*, *going to*, is, *what is and who is*; the optionality of *that* reduction; dative alternation for the verbs *donate*, *suggest*, *create*, *pour*; transitivity for the verbs, *disappear*, *vanish*, *arrive*, *come, fall*. Table 1 shows example quadruplets for each construction (ordered by learnability) showing sentences for (a) the restricted form, (b) the un-restricted form and (c–d) corresponding sentences for which restricted and un-restricted forms are both grammatical (see Hsu and Chater (2010) for details). There was a large spread in learnability predictions. Some constructions appeared learnable within a few years whereas others required years beyond human life spans. Hsu and Chater (2010) compared predicted MDL learnability with child grammar judgments from previous experimental work (Ambridge, Pine, Rowland, & Young, 2008; Theakston, 2004). It was found that child grammar judgments were better correlated with MDL learnability than with frequency counts. However, the comparison with child judgements was limited to a handful of constructions. Here, we wish to test learnability predictions for the full range of constructions analysed in Hsu and Chater (2010). To do so we hypothesise that learnability should also correlate with *adult* grammaticality judgments: In particular, the easier a construction is to learn, the greater the relative difference should be between judgments of the ungrammatical vs. grammatical uses of the construction. It is important to measure relative grammaticality because semantic and syntactic contexts may affect perceived grammaticality (e.g., the context in which we use the verb *disappear* may appear more grammatical than the context in which we use the verb *arrive*). Here we make the first-order assumption that contributions to grammaticality perception add linearly. Therefore, in order to measure knowledge of the restriction rule without the effects of syntactic or semantic context, we will subtract ratings of the grammatical form from ratings of the ungrammatical form (i.e., $a–b$). Furthermore, lexical or syntactic differences between the un-restricted vs. restricted forms of a construction (a vs. b) may also influence grammaticality, independent of the restriction rule. For example, contractions or the transitive usage may be perceived as inherently less grammatical than un-contracted words or the intransitive usage. Thus, we also normalize our measure by subtracting out the grammaticality differences between the related pairs of sentences for which both forms are grammatical (e.g., sentences c and d). This allows us to take into account grammaticality differences that may be due to variations in sentence form (e.g., contraction vs. no-contraction, transitive vs. intransitive) which are not related to knowledge of the restriction rule that we are testing. Thus our measure of relative grammaticality will be the differences between sentences a and b subtracted by differences between sentences c and d, i.e., $(a–b)–(c–d)$. The method of using relative grammar judgments to test linguistic learnability has been previously used in children (Ambridge et al., 2008).

4.1. Model predictions

The most appropriate type of corpus for making learnability predictions is that of child-directed speech, e.g., CHILDES database (Mac Whinney, 1995). However, because many constructions do not occur often enough for

[4] These formal definitions are not directly used in learnability analyses because it is unlikely that first language learners are acquiring grammatical knowledge at this level.

[5] Hsu and Chater (2010) also included analysis of rules concerning the necessary transitivity of the verbs *hit* and *strike* and the dative restriction on *shout* and *whisper*. However, *hit* and *strike* have ambitransitive usage in colloquial speech: In COCA there are 3678 and 1961 intransitive occurrences of *hit* and *strike* respectively, e.g. *The storm hit. Lightning struck.* Also, recent work has suggested that manner-of-speaking verbs such as *shout* and *whisper*, while not traditionally partaking in the dative alternation, actually can alternate (Nikitina & Bresnan, 2009). Thus we did not include these verbs in our experiment.

statistical significance, Hsu and Chater (2010) analysed only four constructions using CHILDES. Therefore, we use model predictions obtained in Hsu and Chater (2010) using the full Corpus of Contemporary American English (COCA), containing 385 million words (90% written, 10% spoken), a reasonable representation of the distributional language information that native English language speakers receive. Learnability results using the British National Corpus were similar to that from COCA (Hsu & Chater, 2010). Fig. 2a shows the estimated number years, N_{years}, required to learn the 15 constructions (Hsu & Chater, 2010). N_{years} was calculated as O_{needed}/O_{year} where O_{needed} is the number of occurrences needed for learning under MDL and O_{year} is the number of occurrences per year, estimated from COCA. We quantify learnability as $\log(1/N_{years})$ (Fig. 2c). This puts O_{year} in the numerator, allowing for direct comparison with the entrenchment hypothesis which compares grammar judgments with occurrence frequency (see Fig. 2b and d and section below). All frequency estimates were estimated by manually counting all relevant sentence frames. Relevant sentence frames were obtained using the search tools offered through Mark Davies' online corpus analysis site (Davies, 2008). The learnability estimates depend on an assumption of the number of total symbols used in a child's original grammar. Here we present results assuming 100,000 total symbols. The relative learnability does not change depending on the assumed number of total symbols. However, the general scale does change, e.g., when assuming 200 total symbols, number of years needed is approximately halved for all constructions. Thus the learnability results of Hsu and Chater (2010) are best interpreted as quantifiers of relative rather than absolute learnability.

4.2. Relation to entrenchment

MDL learnability bears some relation to entrenchment theory (Brooks, Tomasello, Dodson, & Lewis, 1999). According to entrenchment theory, the likelihood of a child over-generalizing a construction with a restriction is inversely related to the occurrence frequency of the construction's un-restricted form, i.e., sentences *b* in Table 1. Previous work has shown that construction occurrence frequency can often predict children's over-generalization errors (Ambridge et al., 2008; Theakston, 2004). However, entrenchment theory lacks a computational explanation: it posits that children avoid over-generalizing commonly heard constructions without offering a feasible method for how this is learned. In contrast, MDL provides a principled account of how retreat from over-generalization may be computed through a cognition-general mechanism for probabilistic learning. Learnability and entrenchment predictions are often related because high construction occurrence frequencies do aid learnability. Thus, learnability may provide a principled explanation for the success of entrenchment theory.

The predictions of entrenchment and learnability will also sometimes differ because of the additional factors that affect learnability (see Fig. 2). These other factors

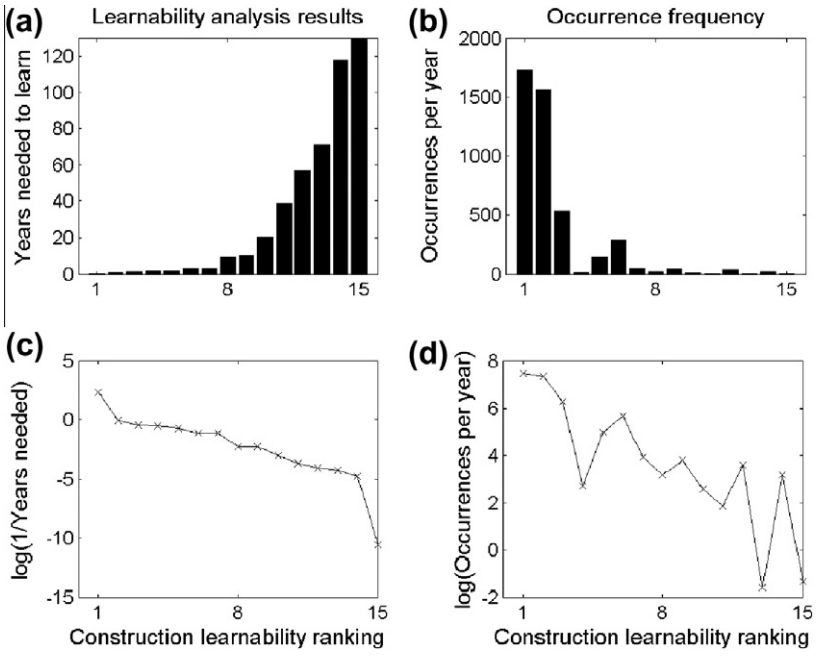

Fig. 2. *Learnability vs. occurrence frequency*: (a) Estimated years required to learn construction. (b) number of occurrences per year (estimated from COCA). (c) Learnability predictions: log(1/years needed). (d) Entrenchment predictions: log(number of occurrences per year). Results summarized from (Hsu & Chater, 2010). The constructions are sorted according to learnability: (1) *is*, (2) *arrive* (3) *come*, (4) *donate*, (5) *fall*, (6) *disappear*, (7) *what is*, (8) *pour*, (9) *vanish*, (10) *create*, (11) *who is*, (12) *going to*, (13) *suggest*, (14) *that*, (15) ∗*want to.* ∗Predicted years for learning *want to* is 3800 years. Yearly occurrence frequencies are estimated by assuming a child hears ∼6 million words per year (Hart & Risley, 1995) and dividing this number by the 385 million words that COCA contains. Note that while more learnable constructions do tend to have higher occurrence frequencies, the two are not completely correlated.

include the complexity of the grammatical rule to be learned, and the relative occurrence probabilities of restricted vs. un-restricted forms in related constructions, e.g., sentences *c* vs. *d* in Table 1. Constructions with significant differences between learnability and entrenchment predictions include *donate*, *suggest* and *that*: Learnability of *suggest* and *donate* are relatively higher and learnability of *that* is relatively lower than mere occurrence frequencies would predict. High encoding savings contribute to the learnability of *donate* and *suggest* because the direct dative is more common than the prepositional dative in the similar verbs *give* and *tell*. This makes the absence of direct dative forms of *donate* and *suggest* more surprising, and easier to learn under MDL than occurrence frequency would predict. The lower learnability of *that* comes from the complexity of the rule governing optionality of *that* insertions, not accounted for by frequency. In order to show the relationship between learnability and entrenchment results, we will also compare adult grammar judgments to construction occurrence frequencies (estimated from COCA).

5. Experimental method

5.1. Participants

200 native English speakers (50 per condition) were recruited for an online grammar judgment study (age range: 15–96 years, mean: 27 years). The majority (90%) of participants learned English in the United States. Other countries included the UK (6%). The remaining were from Australia, Canada and Ireland.

5.2. Procedure

We collected ratings for two sets of quadruplets consisting of sentence types a–d for each of our fifteen constructions. This means there were two examples of each construction analyzed (see Table 1). We used a Latin square design with four conditions. Each participant saw only one sentence type for each quadruplet, i.e., condition 1 participants saw sentences 1a, 2b, 3c, 4d, 5a, etc. In order to balance grammatical vs. ungrammatical sentences in all conditions, we included a 16th quadruplet not analysed in our study (see Table 1). There were 128 sentences total (each participant saw 32 sentences). Trials were block randomized for type (a–d) as well as over-all order of constructions. All constructions from one set were presented before the other set. The set that was presented first was randomized among participants. Sentences were presented visually without audio. Participants were told to assess grammaticality by "what sounds natural" and were encouraged to say the sentences out loud. Participants rated sentences on a scale from 1 to 5: (1) Sounds completely fine (Definitely grammatical), (2) Probably grammatical (Sounds mostly fine), (3) Sounds barely passable (Neutral), (4) Sounds kind of odd (probably ungrammatical), (5) Sounds extremely odd (Definitely ungrammatical).

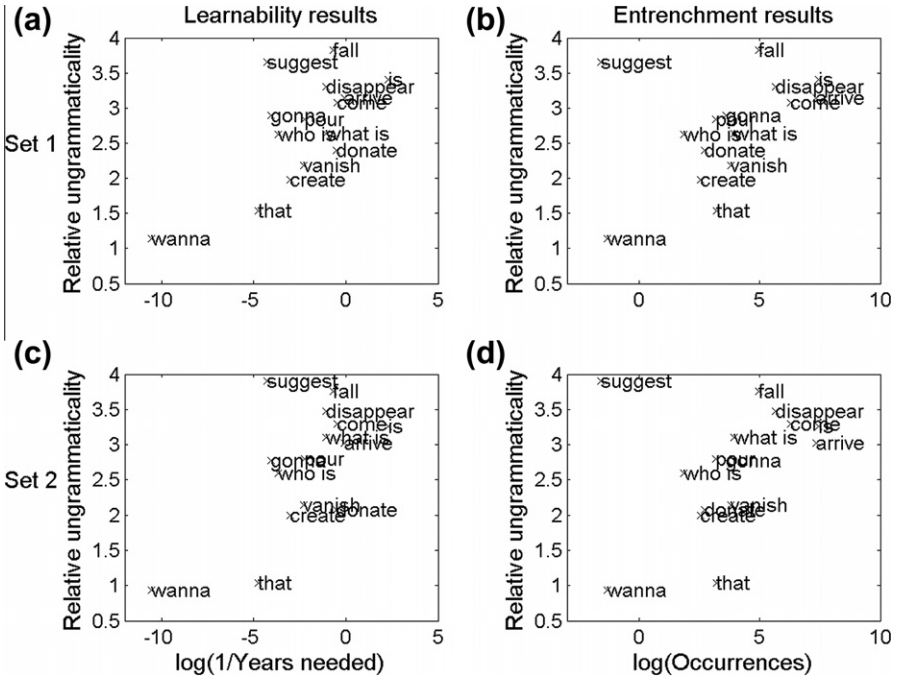

Fig. 3. *Human grammar judgments vs. learnability and occurrence frequency:* (a) Relative grammaticality vs. learnability for Sentence Set 1 ($r = 0.67$; $p = 0.006$). Relative grammaticality is ratings of sentences (a–b)–(c–d) from Table 1. Learnability is log of the inverse of the number of estimated years needed to learn the construction. (b) Relative grammaticality vs. log occurrence frequency ($r = 0.43$; $p = 0.11$). Occurrences is the number of occurrences per year (estimated from COCA) of the un-restricted form of the construction to be learned (sentence *b* in Table 1). (c) Learnability vs. relative gramamticality for Sentence Set 2 ($r = 0.63$; $p = 0.012$). (d) Log occurrence frequency vs. relative grammaticality for Sentence Set 2 ($r = 0.38$; $p = 0.16$).

6. Results

Results show strong correlations between averaged relative grammaticality and MDL learnability for both sets of sentences, $r = 0.67$; $p = 0.006$ and $r = 0.63$; $p = 0.012$, (see Fig. 3a and c). In contrast, grammaticality and log occurrence frequency are not significantly correlated, $r = 0.43$; $p = 0.11$and $r = 0.38$; $p = 0.16$ (see Fig. 3b and d). In particular, learnability is substantially more correlated with grammaticality judgments for restrictions on *suggest* and *that*.

7. Summary and conclusions

This work helps evaluate how much of first language is probabilistically acquired from exposure. We show that, despite putative "logical problems of language acquisition," *any* language generated from any computable generative probability distribution (including any grammars proposed in linguistics) can be precisely identified, given a sufficiently large i.i.d. sample. Our Universal Induction Algorithm embodies no language-specific knowledge, and therefore indicates that language acquisition is possible in principle, given sufficiently large amounts of positive data, and sufficient computing power.

How practically learnable are the types of linguistic patterns that have been often cited as challenges for learnability? To address this, we described a recently formulated framework which allows probabilistic learnability to be quantified. Together, these analyses contribute to a substantial body of work showing that probabilistic language learning *is theoretically and computationally possible*.

Does such probabilistic learning occur in practice? Here we propose that if language is probabilistically acquired, then this should leave traces in adult grammar judgments. MDL learnability assumes that a grammar is learned in an absolute sense: once a grammar is chosen under MDL, that is the one used and there is no gradation of knowledge. However, here we conjecture that learnability should not only correlate with how much data is required to learn a linguistic rule, but also the degree of confidence in that knowledge. Experimental results showed that predicted learnability correlates well with relative grammar judgments for the 15 constructions analyzed, chosen as controversial cases from the literature. Our experimental results thus support the possibility that many linguistic constructions that have been argued to be innate may instead be acquired by probabilistic learning.

Acknowledgements

This work was supported by grant number RES-000-22-3275 (to Hsu and Chater) from the Economics and Social Research Council, and by the ESRC Centre for Economic Learning and Social Evolution (ELSE), UCL. The work of P.M.B. Vitányi was supported in part by the BSIK Project BRICKS of the Dutch government and NWO. Vitányi, Chater and Hsu were all partially supported by the EU NoE PASCAL 2 (Pattern Analysis, Statistical Modeling, and Computational Learning).

Appendix A. Proof of the Computable Probability Identification Theorem

A function is *computable*, if there is a Turing machine (or equivalent) that maps the arguments to the values. Here we consider only probability mass functions with rational arguments. Such a function assigns a probability to each of its arguments (which are countable). By contrast, in statistics and machine learning, many probability distributions have arguments with continuous values. For example, take normal distributions over the reals. Mean and standard deviation can be any real number.

The restriction to computable probability mass functions is both cognitively realistic (if we assume a language is generated by a computable process, a standard assumption in cognitive science) and dramatically simplifies the problem of language identification (for related discussion in a different context, see Cover, 1973).

If a computable function has as values pairs of nonnegative integers, such as (a, b), we can interpret this value as the rational a/b. A real-valued function $f(x)$ with x rational is *semi-computable from below* if it is defined by a rational-valued computable function $\phi(x, k)$ with x a rational number and k a nonnegative integer such that $\phi(x, k+1) \geqslant \phi(x, k)$ for every k and $\lim_{k\to\infty} \phi(x, k) = f(x)$. This means that f can be computably approximated arbitrary closely from below (see Li & Vitanyi, 2008, p. 35), as k increases.

Consider a subclass of functions which are semi-computable from below. A function f is a semiprobability mass function if $\Sigma_x f(x) \leqslant 1$ and a probability mass function if $\Sigma_x f(x) = 1$. We write '$p(x)$' for '$f(x)$' if the function is a semiprobability mass function, and we consider semiprobability mass functions which are semi-computable from below.

We use the following general strategy in the proof. First we enumerate all semiprobabilities which are semi-computable from below $q_1, q_2, \ldots$. That is, we linearly order them with a least element. Note that a computable probability mass function is a fortiori a semiprobability mass function which is semi-computable from below. Thus, the given enumeration contains every computable probability mass function. Second, therefore our target probability p occurs in this list. In fact, we can show it occurs multiple times. We consider the least index k such that $p = q_k$. Third, we give an algorithm that outputs at every step the present candidate which is the least indexed semiprobability mass function which is semi-computable from below, and still compatible with the data. Meanwhile the algorithm discredits candidates that are incompatible with the data seen so far according to a criterion derived from the Strong Law of Large Numbers. Every candidate $q_1, \ldots, q_{k-1}$ gets discredited eventually, each at a certain time say $t_1, \ldots t_{k-1}$ – Eventually, for $t > \max\{t_1, \ldots, t_{k-1}\}$ the algorithm will output at every step the least indexed nondiscredited distribution, that is, $q_k = p$. We now show how this proof strategy can be carried out in more detail.

It is possible to enumerate all and only the semiprobability mass functions that are semi-computable from be-

low, by fixing an effective enumeration of all Turing machines in a fixed description syntax. Now it is possible to change every Turing machine description in the list into one that computes a semiprobability mass function that is computable from below, as described in the proof of Theorem 4.3.1 in Li and Vitanyi (1997, 2008). The list contains all and only semiprobability mass functions that are semi-computable from below.

Every probability mass function is a semiprobability mass function, and every computable probability mass function is semi-computable from below. Therefore, every computable probability mass function is in the list (indeed, each will appear infinitely often).

Definition 1. In probability theory the statement almost surely means "with probability one."

Let us illustrate this notion for infinite objects. It is possible that a fair coin (a $(\frac{1}{2},\frac{1}{2})$ Bernoulli process) generates an infinite sequence 0, 0,... even though the probability of 1 is $\frac{1}{2}$. Consider the property that the relative frequency of 1s goes to the limit $\frac{1}{2}$. The uniform measure of the set of those infinite sequences is one. Hence, the probability that an infinite sequence is of that type is one, even though there are infinite sequences (like in the example above) that are not of that type. Thus, "almost surely" for infinite objects may not mean "with certainty."

Theorem 1. (Computable Probability Identification Theorem). *Let L be a language $L = \{a_1, a_2, \ldots\}$ (a countably finite or infinite set), and p a computable probability mass function such that the probability of a_i is $p(a_i)$ for $i = 1,2,\ldots$. Let the mean of p be finite* (2). *Then, p can almost surely be computed by an algorithm that takes as input an infinite sequence $x_1, x_2, \ldots$ of elements of L drawn i.i.d. according to p.*

Proof. Our data is, by assumption, i.i.d. drawn from L according to a computable probability mass function p. Formally, the data $x_1, x_2, \ldots$ are generated by a sequence of random variables $X_1, X_2, \ldots$, each a copy of a single random variable X with probability mass function $P(X = x) = p(x)$ $(x \in L)$. We assume that the mean of p exists, as noted in the statement of Theorem 1. By the preceding arguments, we can effectively enumerate the semiprobability mass functions that are computable from below as

$$Q = q_1, q_2, \ldots$$

Definition 2. Define k to be the least integer such that $p = q_k$.

We now turn to a probabilistic law that makes it possible to compute index k almost surely given data $x_1, x_2, \ldots$. The strong law of large numbers states that if we perform the same experiment a large number of times, then almost surely the average of the results goes to the expected value.

Again, note that it is possible that a fair coin generates an infinite sequence 0, 0,... even though the probability of 1 is $\frac{1}{2}$. For this particular sequence the inequality (1) below does not hold. Hence, the Strong Law of Large Numbers holds "almost surely" and cannot hold "with certainty."

Let $\#a(x_1, x_2, \ldots, x_n)$ be the number of elements in $x_1, x_2, \ldots, x_n$ equal $a (a \in L)$. Consider some $x \in L$. Then, we can consider a Bernoulli process $(q, 1-q)$ where $q = p(x)$ and $1 - q = \sum_{y \in L - \{x\}} p(y)$. Then for every pair $\in, \delta$, there is an N such that for every $r > 0$, all $r + 1$ inequalities:

$$\mid p(x) - \frac{\#x(x_1, x_2, \ldots, x_n)}{n} \mid \leqslant \epsilon, \tag{1}$$

with $n = N, N+1, \ldots, N + r$ will be satisfied with probability at least $1 - \delta$ (Feller, 1968, p. 258 ff). That is, we can say, informally, that with overwhelming probability the left hand part of (1) remains small for all $n \geqslant N$. This holds since our sequence of variables $X_1, X_2, \ldots$ satisfies Kolmogorov's criterion that

$$\sum_i \frac{(\sigma_i)^2}{i^2} < \infty,$$

where $(\sigma_i)^2$ is the variance of X_i in the sequence of mutually independent random variables $X_1, X_2, \ldots$. Since all X_i's are copies of a single X, all X_i's have a common distribution p and mean μ_p. If $\mu_p < \infty$ as assumed, then $\sum_i (\sigma_i)^2 / i^2 < \infty$. This is proved in the (proof of) the theorem on page 260 in Feller (1968). To apply the Strong Law in this case, it thus suffices that $\mu_p < \infty$. If we order the elements of L length-increasing lexicographic, and $i(x)$ is the index of x in the ordered L, then we require that

$$\mu_p = \sum_{x \in L} i(x) p(x) < \infty. \tag{2}$$

Our current guess concerning the language is the earliest element in the list Q of semiprobability mass functions which are semi-computable from below and are not yet ruled out by the data. Since the elements of Q are semi-computable from below, if for some i at some step t we have $q_i^t(x) > p(x)$ then we can rule out q_i (where $q_i^t(x)$ is the semi-computable function approximating $q_i(x)$ from below, at step t). These hypotheses can be permanently eliminated. On the other hand, there may be other hypotheses, q_j, for which $q_j^t(x) < p(x)$ and at some later step $t' > t$ we have $q_j^{t'}(x) \geqslant p(x)$, because lower semi-computable functions may increase (though not decrease) as the computation proceeds. This means that our guess of the earliest $q \epsilon Q$ that is actually p may change with the number of steps t. Thus, the candidates guessed may change from earlier on in the list Q to later into earlier. However, eventually we will identify the correct hypothesis p (rather, eliminate all incorrect previous hypotheses).

To see this, we reason as follows. On one hand, there is a real constant $\alpha > 0$ such that for every $i = 1,2,\ldots,k-1$ there is an $a \in L$ such that $\mid q_i(a) - p(a) \mid \geqslant \alpha$ (if not, then $q_i = p$ for some $i < k$). Moreover, there is a timestep T such that for every $i = 1, 2, \ldots, k-1$ and all $t \geqslant T$ we have $q_i(a) - q_i^t(a) \leq \alpha/2$ and $\mid q_i^t(a) - p(a) \mid \leqslant \alpha/2$. Therefore, $|q_i^t(a) - \#a(x_1, x_2, \ldots, x_n)| > \epsilon$ for $t \geqslant T$ and n large enough for all $\epsilon \leqslant \alpha/4$. That is, (1) does not hold for $q_1, q_2, \ldots, q_{k-1}$.

On the other hand, for every $\epsilon > 0$ there is an integer n_ϵ such that for all $n \geqslant n_\epsilon$ the following holds. There is an integer h and timestep T' such that $1 - \sum_{i=1}^{h} p(a_h) \leqslant \epsilon/2$ and for every $t \geqslant T'$ and every $1 \leqslant i \leqslant h$ we *have* $p(a_i) - p^t(a_i) \leqslant \epsilon/2$ and $\mid p(a_i) - \#a_i(x_1, x_2, \ldots, x_n \mid \leqslant \epsilon/2$.

(The condition means that each of $p(a_h), p(a_{h+1}), \ldots$ is at most $\epsilon/2$, and for an infinite language L we have $h \to \infty$ with $\epsilon \to 0$). Altogether,$|\ p(a_i) - \#a_i(x_1, x_2, \ldots, x_n)\ | \leqslant \epsilon$ for every $\epsilon > 0$ and $1 \leqslant i \leqslant h$ with h as above. That is, (1) holds for $p = q_k$.

Thus, the (finite number of) incorrect hypotheses earlier in the list from the correct hypothesis will eventually converge closely enough to their true (incorrect) probability estimate to be eliminated by reference to the strong law of large numbers. It is entirely possible that the true hypothesis may, early in the computation, be provisionally rejected (because early in the computation, the approximation is too poor); and it may successively be proposed and rejected a finite number of times during the computation. Eventually, however, as the true hypothesis satisfies (1) it will never be eliminated, however much data is obtained and however long the computation proceeds. Thus the true probability distribution is identified. Moreover, it can be shown that this process can be carried out by a concrete algorithm (Chater & Vitányi, in preparation). □

References

Ambridge, B., Pine, J., Rowland, C., & Young, C. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition, 106*, 87–129.

Angluin, D. (1988). *Identifying languages from stochastic examples. Technical Report YALEU/DCS/RR-614*. Yale University, Department of Computer Science, New Haven, CT.

Baker, C. L., & McCarthy, J. J. (1981). *The logical problem of language acquisition*. Cambridge, Mass: MIT Press.

Bowerman, M. (1988). The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In J. Hawkins (Ed.), *Explaining language universals* (pp. 73–101). Oxford: Blackwell.

Brooks, P., Tomasello, M., Dodson, K., & Lewis, L. (1999). Young children's overgeneralizations with fixed transitivity verbs. *Child Development, 70*, 1325–1337.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review, 103*, 566–581.

Chater, N., & Vitányi, P. M. B. (2007). Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology, 51*, 135–163.

Chater, N., & Vitányi, P. M. B. (in preparation). Computable probability identification.

Chomsky, N. (1975). *The logical structure of linguistic theory*. London: Plenum Press.

Clark, A., & Eyraud, R. (2007). Polynomial identification in the limit of substitutable context-free languages. *Journal of Machine Learning Research, 8*, 1725–1745.

Cover, T. M. (1973). On the determination of the irrationality of the mean of a random variable. *Annals of Statistics, 1*, 862–871.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences, 14*, 597–612.

Davies, M. (2008). The corpus of contemporary American English (COCA): 385 million words, 1990-present. Corpus of Contemporary American English. <http://www.americancorpus.org>.

Dowman, M. (in preparation). Minimum description length as a solution to the problem of generalization in syntactic theory. *Machine Learning and Language*.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407*, 630–633.

Feldman, J. A., Gips, J., Horning, J. J., & Reder, S. (1969). *Grammatical complexity and inference. (Rep. No. CS 125)*. Stanford University.

Feller, W. (1968). *An introduction to probability theory and its applications* (Vol. 1, 3rd ed.). New York: Wiley.

Foraker, S., Regier, T., Khetarpal, N., Perfors, A., & Tenenbaum, J. B. (2009). Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science, 33*, 300.

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 16*, 447–474.

Griffiths, T., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biasesG. *Trends in Cognitive Sciences, 14*, 357–364.

Grünwald, P. (1994). A minimum description length approach to grammar inference. In S. Scheler, Wernter, & E. Rilof (Eds.), *Connectionist, statistical and symbolic approaches to learning for natural language* (pp. 203–216). Berlin: Springer Verlag.

Hart, B., & Risley, J. (1995). *Meaningful differences in the everyday experience of young american children*. Baltimore, Maryland: Brookes Publishing.

Horning, J. J. (1969). *A study of grammatical inference*. Stanford University.

Hornstein, N., & Lightfoot, D. W. (1981). *Explanation in linguistics: The logical problem of language acquisition*. London: Longman.

Hsu, A., & Chater, N. (2010). The logical problem of language acquisition goes probabilistic: No negative evidence as a window on language acquisition. *Cognitive Science, 34*, 972–1016.

Hsu, A., & Griffiths, T. (2009). Differencial use of implicit negative evidence in generative and discriminative language learning. *Neural Information Processing Systems, 22*.

Li, M., & Vitányi, P. M. B. (1997). *An introduction to Kolmogorov complexity theory and its applications* (2nd ed.). New York: Springer.

Li, M., & Vitányi, P. M. B. (2008). *An introduction to Kolmogorov complexity theory and its applications* (3rd ed.). New York: Springer.

Mac Whinney, B. (1995). *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

Nikitina, T., & Bresnan, J. (2009). The Gradience of the Dative Alternation. In L. Uyechi & L. H. Wee (Eds.), *Reality exploration and discovery: Pattern interaction in language and life (pp. 161–184)*. Stanford: CSLI Publications.

Nowak, M., Komarova, N., & Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature, 417*, 611–617.

Osherson, D., Stob, M., & Weinstein, S. (1985). *Systems that learn*. Cambridge, MA: MIT Press.

Perfors, A., Regier, T., & Tenenbaum, J. B. (2006). Poverty of the stimulus? A rational approach. *Proceedings of the Twenty-eighth Annual Conference of the Cognitive Science Society, 663*, 668.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Regier, T., & Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. *Cognition, 93*, 147–155.

Theakston, A. (2004). The role of entrenchment in children's and adults' performance on grammaticality judgment tasks. *Cognitive Development, 19*, 15–34.

Vitányi, P. M. B., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory, IT, 46*, 446–464.